Virtual Institute of Microbial Stress and Survival
DOE Genomes To Life Project
Progress Report: July, 2003

I. Overview

The objective of this monthly progress report is to provide an update of the technical and administrative actions from the previous month as well as forecast upcoming progress for the VIMSS Genomes to Life Project. I want to remind everyone how important to make sure everyone is communicating. The discussion boards (http://genomics.lbl.gov/~aparkin/discus) provide a forum for people to ask questions about direction of the project, priorities, and technical issues that can be read and answered by the entire group. I know email is often the most efficient means but it does privatize some of the important communications. Also, posting project data and information to BioFiles (https://tayma.lbl.gov/perl/biofiles) is EXTREMELY important. We are in the process of adding user help files to BioFiles – if you have user questions, please contact Keith Keller (tel: 510.495.2766 or email: kkeller@lbl.gov). This is the best metric I can give to the DOE leadership that we are making progress aside from the VIMSS website. Please make us and yourselves visible by donating data and information to the website.

II. Applied Environmental Microbiology Core

LBNL

SR-FTIR. We continue to modify the existing SR-FTIR spectromicroscopy apparatus to study *Desulfovibrio vulgaris* under anaerobic conditions. Continued to develop protocols for producing *Desulfovibrio vulgaris* biofilm on reference surface and on different types of mineral surfaces. Continued to establish the IR spectral baseline of *D. vulgaris* biofilm under anaerobic conditions. Validated IR results by live-dead stain fluorescence microscopy. We succeeded in (1) developing one of a series of protocols to produce *D. vulgaris* biofilm on a reference surface, (2) obtaining the IR spectral baseline of *D. vulgaris* biofilm under anaerobic conditions, and (3) confirming the IR results by means of live-dead stain microscopy. We postponed the study using medium without the addition of Fe as a growth indicator.

This month the work in the biomass production lab was focused on measuring growth curves in both air and N_2 -sparged reactor systems. The overall objective of this first experiment was to (1) define baseline growth patterns that occur over 12h or one generation time as measured by the techniques being developed, (2) determine logistical constraints for sampling, shipping etc. (3) establish QA/QC verification of assay protocols and (4) develop functional design criteria for future experiments. The parameters for the experiments were 10% inoculums of primary culture, 30°C incubation, shake at 100 rpm, and BAARs modified medium (-Fe). Six bottles were

prepared; two air-sparged at a rate of 57 ml/min for a final concentration of O_2 of 8 mg/l, two N_2-sparged at 57 ml/min, and two non-sparged anaerobic control (see figures 1 and 2). Cell purity was determined by colony morphology (plates), cell morphology (microscopy), phenotypic characterization (PLFA), DNA characterization (RFLP).

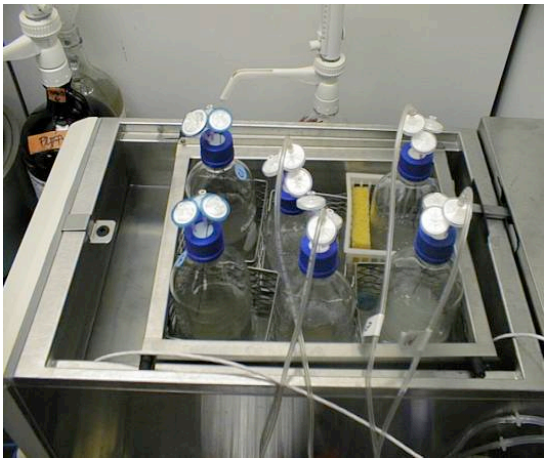


Figure 1. Apparatus for cell growth apparatus

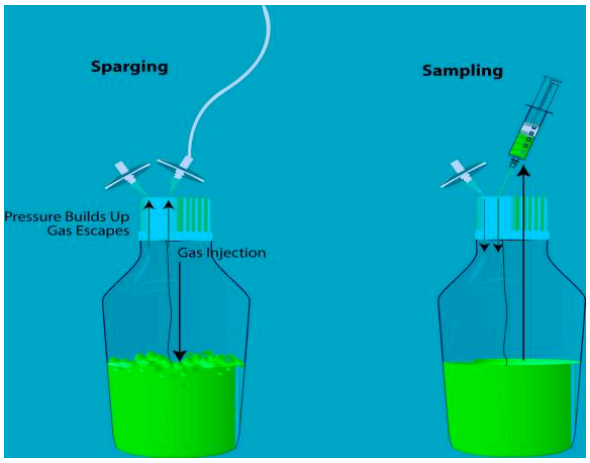


Figure 2. Schematic of growth

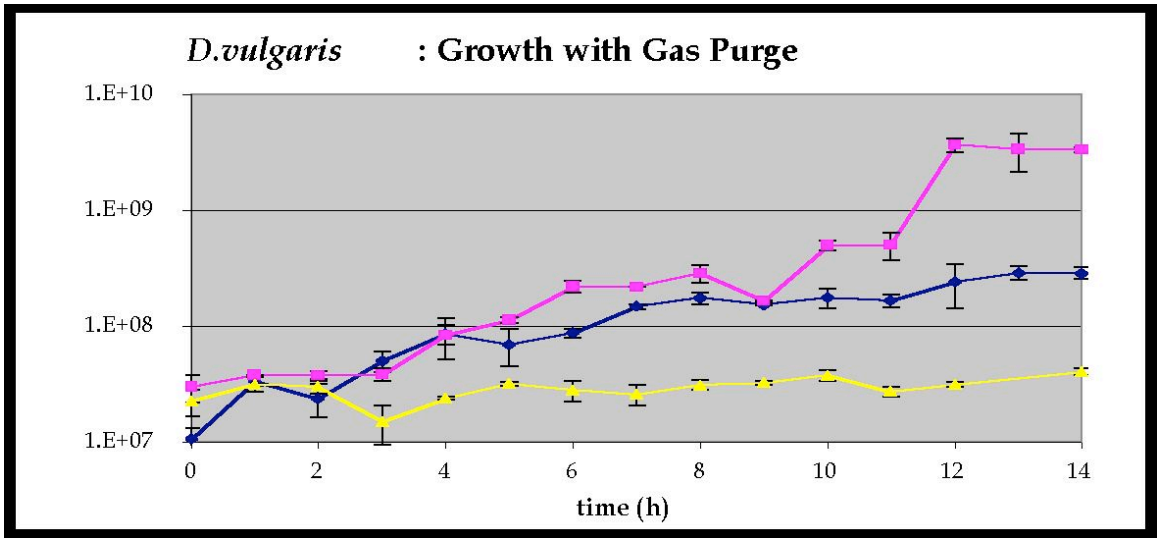


Figure 3. AODC (Cell counts) from growth curve.

At 10% inoculum of log phase primary culture, experimental log phase occurs in the range of 3-13 h with a generation time of 1.5-3.0 h. The air-sparged cultures showed no significant changes in any of the parameters over the same time periods; however, cells were viable. Differences in growth patterns of sparged and closed systems suggest that H_2S may inhibit growth in the absence of Fe. OD, AODC, and protein assays gave comparable results for all experiments, though protein assays did not show as great a

difference between sparged and non-sparged incubation. AODC data is shown in Figure 3. Phenotypic characterization of growth curve cells by PLFA showed significant lipid changes in the N_2 sparged system (Figure 4). Measurable precipitate accumulates as a by-product of growth in both Fe (+) and Fe (-) medium. It is not intact biomass as determined by DNA staining. The data suggest survivability of *D. vulgaris* in the presence of oxygen. The next step will be to demonstrate recovery and growth of an aerated culture as the oxygen concentration recedes. In further studies, dissolved oxygen in the medium will be measured with a fiber optic oxygen sensor to determine the concentration of oxygen in the system that inhibits growth of *D. vulgaris.*

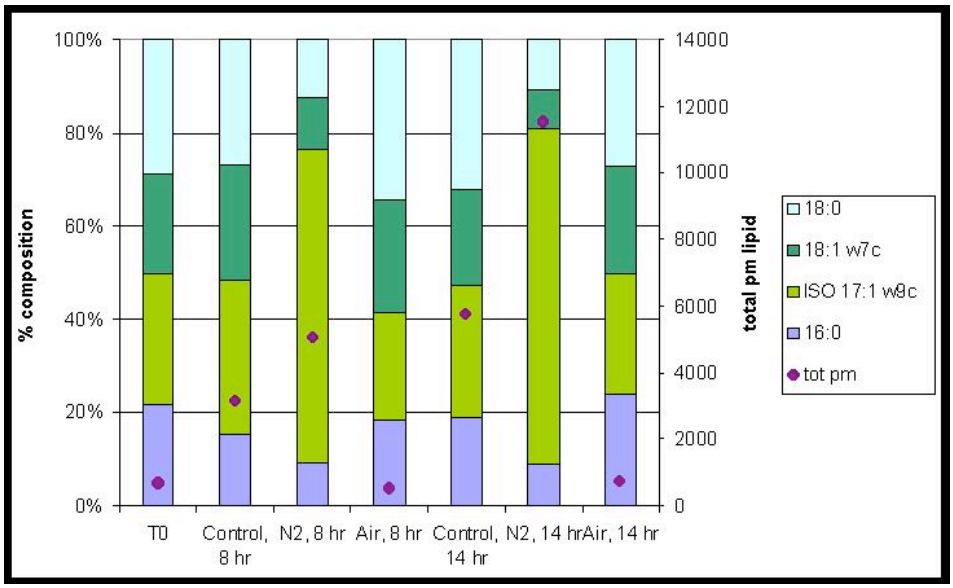


Figure 4. Signature lipids (PLFA) from growth curve.

Figure 5. *D. vulgaris,* modified BAARS (-Fe), no sparge.

University of Washington

MPN enrichments of new sediment sample FWB203-03d 04 from FRC area 2 which have been started on B3 media supplemented with one of the following substrates: 1) lactate, 2) acetate, 3) propionate, 4) pyruvate, 5) ethanol and 6) hydrogen with carbon dioxide demonstrated growth on all substrates in all replicates of a 10-1 to 10-4 dilution series. Sulfide production was detected in the enrichments supplemented with each growth substrate tested.

Single colonies obtained earlier from ground water TPB-25Y of FRC area 2 enrichment supplemented with lactate plus ethanol and demonstrating sulfate reducing activity were transferred to liquid B3 medium supplemented with lactate plus ethanol but have yet to show visible growth.

We set up new co-cultures of *Desulfovibrio* strains (Hildenborough and PT2) with *Methanococcus maripaludis* in a modified McC medium containing a reduced concentration of NaCl (2g/L). Although initial growth has been slow, methane is being produced (up to 1% methane in the head space with *D. vulgaris* Hildenborough and up to 1.9% for with *Desulfovibrio* species strain PT2) after 48hr incubation.

The FairMenTec Bioreactor and fixed bed bioreactor systems have been relocated to a remodeled room, reassembled and prepared for experiments.

Near-complete 16S rRNA genes of 16 Lake Depue isolates were cloned and sequenced. Seven of were 98% identical to *Acidaminobacter hydrogenoformans* (accession number AF016691). Nucleotide sequences of the nine other isolates demonstrated 99% identity to 16S gene from D. vulgaris Hildenborough.

Immediate future work

- Continue optimizing co-culture growth conditions on modified McC medium.
- Obtain and analyze dsrAB gene sequences of SRB isolated from Lake Depue metal-contaminated sediments.
- Amplify and sequence 16S rRNA and dsrAB genes from FRC isolates.
- The fixed bed reactor will be tested for its ability to maintain prolonged anaerobic conditions inoculated with a pure culture *D. vulgaris* Hildenborough.

Oak Ridge National Laboratory

Anaerobic nitrate-reducing enrichments were initiated with FRC sediment (~ 15 m) with a combination of lactate and ethanol as potential electron donors. A medium that simulates the geochemical conditions of groundwater was used, and the nitrate-reducing cultures provided an acceptable amount of biomass for optimization of protocols. A 10^{-2} dilution was chosen for further work and extraction of high molecular weight (HMW)-DNA. The nitrate-reducing enrichment was predominated (over 50%) by an OTU that had 99% sequence identity with an uncultivated clone from an acid-tolerant epiphytic bacterial community. The predominant population also had 99% sequence identity with an isolate from FRC groundwater that appeared to be a *Pseudomonas* species. The enrichment clone was similar to sequences cloned directly from FRC groundwater that constituted approximately 7 to 10% of the sampled diversity. A second population accounted for approximately 30% of the enrichment community, and had 97% sequence identity with *Citrobacter sedlakii*. The data also suggested that microorganisms that had 96% sequence identity with *Azoarcus eutrophus* comprised 3% of the community. Less than 1% of the community appeared to be a-*Protoebaceria* with approximately 93% sequence identity with *Ochrombactrum intermedia*, and the OTU had between 94% and 96% sequence identity with two bacterial isolates from FRC groundwater. Preliminary results with HMW-DNA from the enrichment indicated that sizable libraries could be produced (>8,000 clones) with an average insert size between 35 to 40 kb when fosmid vectors are used.

Using the same medium and electron donors as above (lactate and ethanol), the same sediment inoculum was used to initiate sulfate-reducing enrichments. The predominant population comprised approximately 25% of the sampled diversity and had 88% sequence identity with *Desulfosporosinus* Blif. Subpopulations that had 95% to 97% sequence identity with *Desulfosporosinus orientis* constituted for an additional 37% of the library. The clone E04-023 constituted just over 10% of the library, and had 98% sequence identity with *Clostridium chromoreductans*. Two other clones that comprised an additional 10% of the library also clustered with *C*. *chromoreductans*. *C*. *chromoreductans* was recently isolated from soil, and has the capacity to reduce Cr(VI). A small fraction of the enrichment community (5%) appeared to have only 87% sequence identity with previously uncultivated clones that originated from a chlorobenzene-degrading community. The construction of large insert libraries from the metagenomic DNA of the sulfate-reducing enrichment is underway.

Based on data to date, microorganisms closely related *Shewanella*, *Geobacter*, or *Desulfovibrio* have not been observed in groundwater communities or enrichment cultures from sediments. Unpublished, preliminary results from other researchers of sediments in Area 3 have detected *Desulfomaculum* as a dominant phylogenetic group, and a recent study of enrichment cultures from a uranium mill tailings site suggested the presence of low G+C Gram-positive microorganisms when the medium closely matched site geochemistry. However, limited sites have been investigated at the FRC. Further

work is needed to identify dominant phylogenetic groups that contain the desired phenotypes and cultivate novel isolates.

Diversa

Progress

- A poster entitled “From Soil to Microbes to Genes” was prepared for the Genomes to Life retreat. Topics covered include large fragment DNA extraction, amplification, library construction, and library screening.

- Genomic DNA amplification reactions continue to be optimized for product size and genome coverage.

- The large insert DNA recovery procedure has been optimized for the large insert FACS biopanning protocol.

Actions

- Methods to concentrate gDNA from extremely dilute noodle extractions are continuing to be explored.

- Diversity indexing will be used to evaluate different amplification methods for small and large fragment genomic DNA.

- High GC random primers have been used to amplify *S. diversa* genomic DNA, which is GC rich (~70%). This DNA will be hybridized to an Affymetrix GeneChip to look at genome coverage.

- Work is ongoing to optimize the large insert FACS biopanning protocol in gel microdroplets. Currently, experiments are in progress to increase the positive hit rate on the FACS, and decrease the background.

III. Functional Genomics Core

Transcriptomics (ORNL)

Objective

- To create and use whole-genome microarrays for *Desulfovibrio* and *Geobacter* for analysis of stress-induced transcriptomes.

Progress since last report

- ArrayOligoSelector software was used to design oligonucleotide probes for *Desulfovibrio vulgaris* and *Geobacter metallireducens* (approximately 70 to 80% coverage for *G. metallireducens*).

- Conditions are being determined first with synthesized oligonucleotides with known mismatches, and the affects on hybridization signal intensity and cross-hybridization determined.

- Preliminary data suggests that 70mer oligoarrays perform well at 45 to 50°C with 50% formamide in the hybridization buffer when compared to an analogous cDNA array.

- Preliminary results also suggest that target DNA concentrations will affect the signal intensities, and further work is needed to determine the effects with genomic DNA and cDNA. Some preliminary results are given below with cDNA from *D*. *vulgaris* cells harvested during logarithmic or stationary phase.

- Probes on the microarray were based upon predicted ORFs for *G. metallireducens*. The array was hybridized with cDNA from *D. vulgaris* cells in the logarithmic phase or stationary phase of growth. Signal intensities differed from spot to spot, but a majority of probes displayed some degree of hybridization at 45°C with 50% formamide. The true extent of cross hybridization is not yet known, but the results suggested that higher hybridization temperatures are required. Interestingly, enzymes indicative of logarithmic growth had significantly increased hybridization signals compared to stationary phase cells. For instance, probes for aconitase, fructose-1,6-bisphosphatase, and glycogen synthase had 33-fold, 175-fold, and 24-fold increases for logarithmic phase cells, respectively. Interestingly, a gene annotated as a lipoprotein that has some homology with an Omp protein associated with the transport of solutes at low concentrations was up-regulated 4-fold in stationary phase cells.

- Another interesting observation struck me after learning that *D. vulgaris* cells appeared to loose flagella after being exposed to oxygen. In relation to the different stress response experiments with *Shewanella oneidensis* using the microarrays, flagellar genes are down-regulated dependent upon the stress. The large flagellar operon is down-regulated in response to salt stress and temperature stress, but in general not oxidative (H_2O_2) stress or pH shifts. Some genes did display down-regulation in response to an alkaline environment. It will be interesting to determine which groups of genes respond to multiple stresses, but not others. Hopefully in the future the computational core will be able to compare the *Shewanella* microarray data from the different stresses to try to identify common and different responses across similar groups of genes (i.e., maybe call them responsulons?).

Future work

- Work continues to optimize hybridization conditions for 70-mer oligonucleotide probes, and untreated glass slides from Corning are being used.

Proteomics (Diversa)

Objective

- The activities are mainly in the area of data analysis.

Progress since last report:

- *Spectra Quality Assessment:* The qualities of the MS/MS spectra acquired from the protein samples vary widely. To reduce the consumption of the computing and human resource, less promising ones should be excluded as early as possible. We've undertaken an effort to formulate a robust indicator of the spectra quality defined as the content of the significant fragment ions.
  1. Spectra, which contain too few peaks, to be viable for peptide identification are removed as a class (sparse).
  2. The baseline of the spectrum is estimated using a local window + outlier rejection algorithm. The result is displayed in **Figure 1**.
  3. The "overall S/N ratio" is calculated as (base peak intensity / max. baseline level).
  4. Spectra with SNR < 4.0 are rejected.

Validation: For a 1D LC-MS/MS dataset acquired from a protein standard. Out of 5235 raw MS/MS, "Sparse" spectra took up 102. The rest of them yielded the overall S/N ratio ranged from 1.5 to 1507. Upon visual examination of the spectra themselves, those with S/N < 4.0 are almost always useless (~650 @ S/N < 4.0). The spectra start to show some discernable feature when S/N > 6.0 (1300 < S/N6.0). They're quite good once the S/N is >8.0 as depicted in

The percent of rejection is ~14%. Using another LC-MS/MS dataset acquired at the end of chromatography gradient, the rejection rate increase to ~ 70% by the same criteria. This represents the worst-case scenario as far as the spectra quality is concerned in practice.

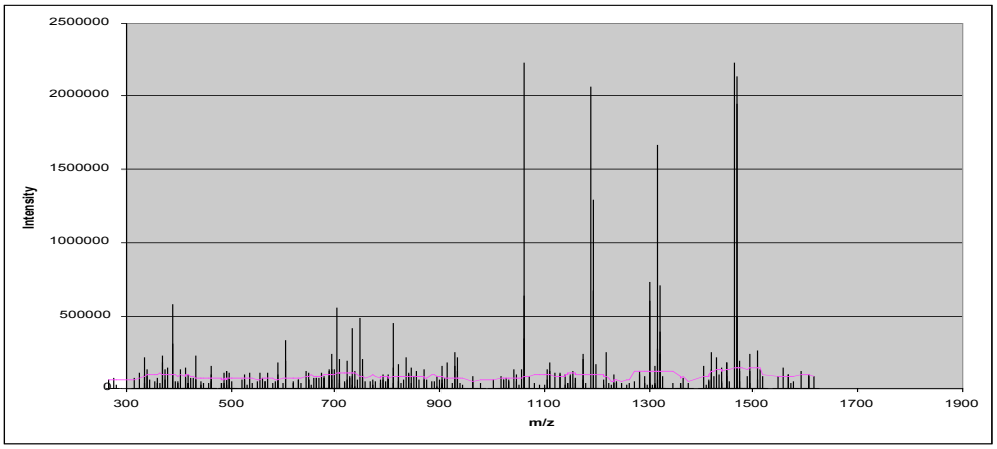


Figure 1. Estimation of the local baseline as indicated by the purple line. The MS/MS spectrum is taken from a sample prepared from the mixtures of known proteins

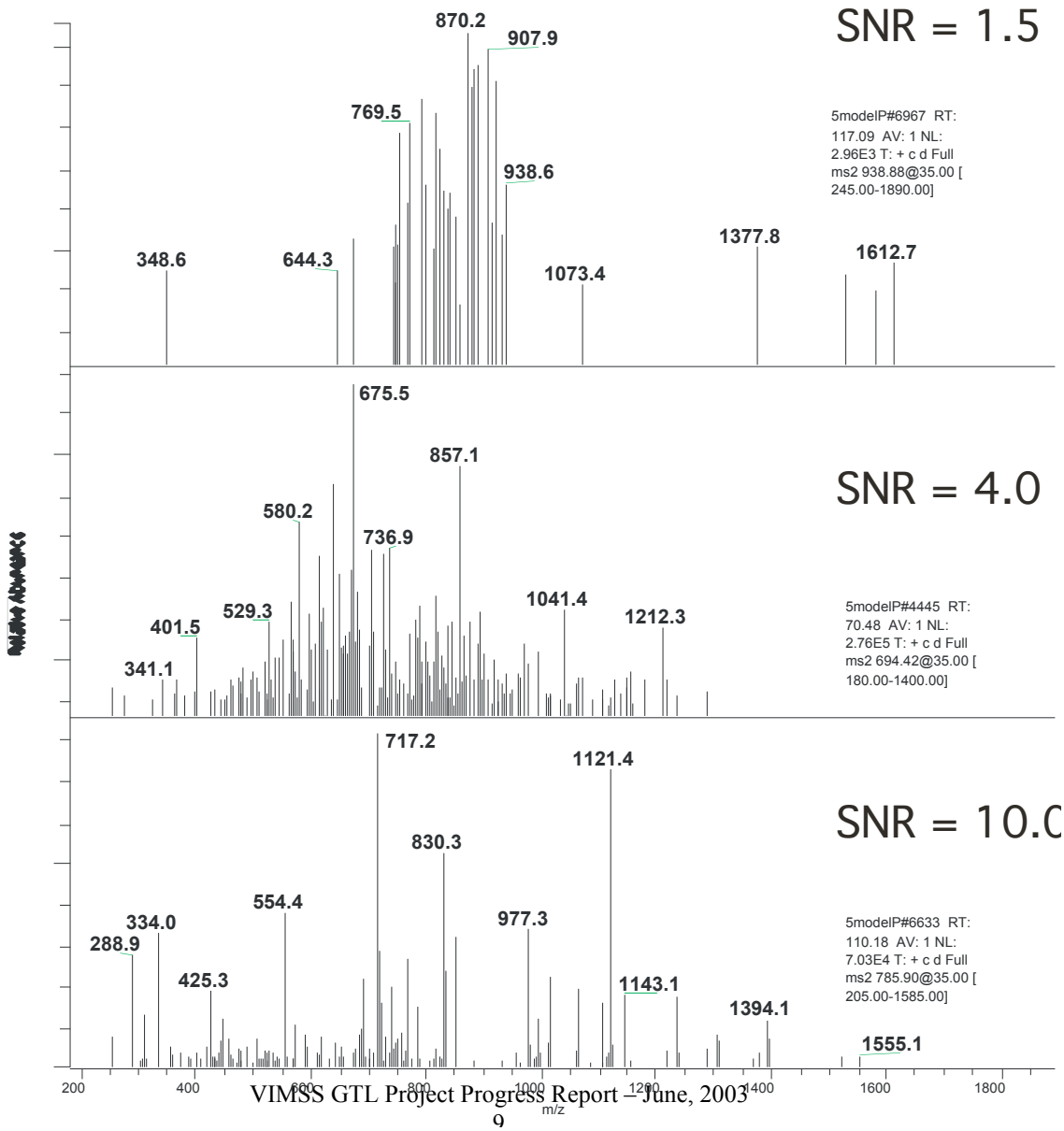
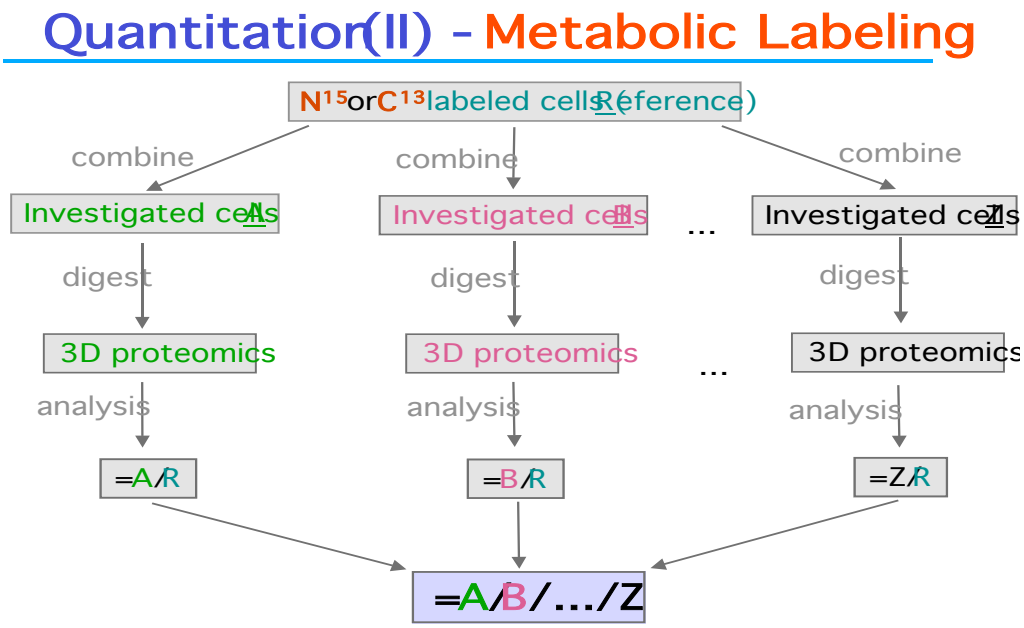


Figure 2. Examples of the MS/MS spectra at different SNR values. Top) poor spectrum; Middle) Borderline; Bottom) Good spectrum.

Method evaluation using metabolic C13 or N15 labeling procedures for differential quantitative analysis.



It is likely to generate C13 labeled *D. vulgaris* strain as a reference for quantitation proteomics. We may compare it to the chemical labeling approach and choose the better one for the stress response study.

Protein complexes (Sandia)

Objective

- To work towards isolation of proteins and complexes involved in heat-shock in *D. vulgaris* and their identification by separation (2D-gel/LC) followed by MS as well as nanoLC/MS/MS of isotope-labeled peptides.

Progress since last report

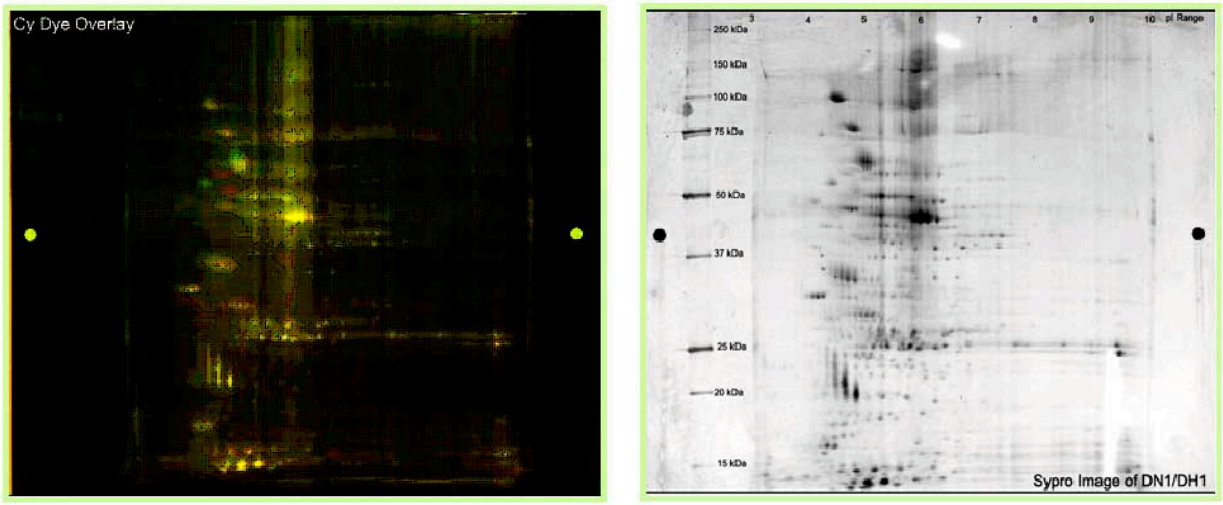
D. vulgaris

- For the heat shock experiment on *D. vulgaris* as described in previous progress reports, DIGE was performed on the control and experimental samples.

- Identified ORFs for 96 spots on DIGE gels (**Figure 3**). 36 spots that displayed high expression levels under both conditions were chosen, 41 spots showed an increase in protein synthesis during heat shock, 12 spots showed a decrease, 6 were positive controls and 1 negative control was used.

- Spots were trypsinized and the mass determined using the Voyager DE PRO MALDI TOF from Applied Biosystems. MASCOT was used to identify ORFs from the *D. vulgaris* proteome database.

- ORF 00281 (Hsp 70) was identified as a target protein with over a 2 fold expression level under the given experimental conditions. Several other proteins were identified. We are currently analyzing the data for the remaining spots.



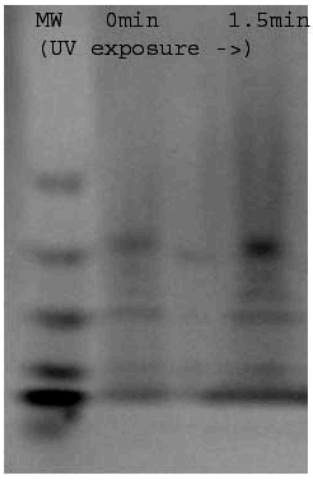
2 Dimensional Gel Electrophoresis of whole cell protein extract from heat shocked (30 min) and non heat shocked (0 min) cells labeled with Cy5 and Cy3 respectively. The overlay identifies over-expressed and repressed proteins. The sypro stain was used to pick spots for identification through MS analysis.

Figure 3.

E. coli

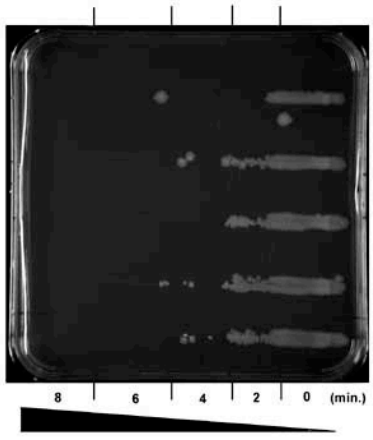
- For the UV stress experiment on E. coli as described in previous progress reports, a co-immunoprecipitation run was performed using Dynal magnetic beads coated with anti-RecA anitbody.

- An SDS-PAGE gel was run (**Figure 4 and 5**) on the "pulled" down complex. The



Silver stained gel showing "pulled" down complex using

Fig 3.



E. Coli K12 cell cultures were plated on Agar

silver stained gel showed the presence of several bands after a 1.5 min exposure.

We are currently in the process of identifying these bands.

Figure 4. **Figure 5.**

Future Work

- Expanding DIGE and ICAT technologies to other stress conditions including exposure to Oxygen, medium basicity and acidity changes.
- Developing improved protocols for identifying protein complexes through magnetic beads and other techniques using anti-HSP70 as the bait protein for heat shock conditions.
- Developing genetic tagging methods for complex identification under stress conditions.

Metabolomics and Proteomics (UCB, LBNL)

Objective

- Optimization of nucleotide separation by hydrophilic interaction chromatography
- Develop a method to separate NAD and NADP
- To establish a knockout strategy by completing the annealing and transformation of the vector into *E. coli*, followed by conjugation into *D. vulgaris*.
- As primers arrive, clone the rest of the histidine sensor kinases out of the genome.
- Complete a preliminary ICAT oxygen stress experiment with samples grown in the Hazen lab.

Progress since last report

- For the nucleotide separation, a condition was obtained to generate symmetric peaks, which is necessary for quantification.
- A hydrophilic interaction chromatographic method was developed to separate NAD and NADP (**Figure 6**).

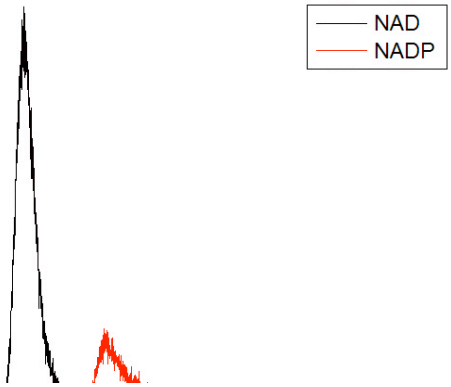


Figure 6. Separation of NAD and NADP. An Amide-80 column was used. The sample was eluted with 20% buffer (10 mM NH4OAc, pH 5.0) for 10 minutes and followed with a linear

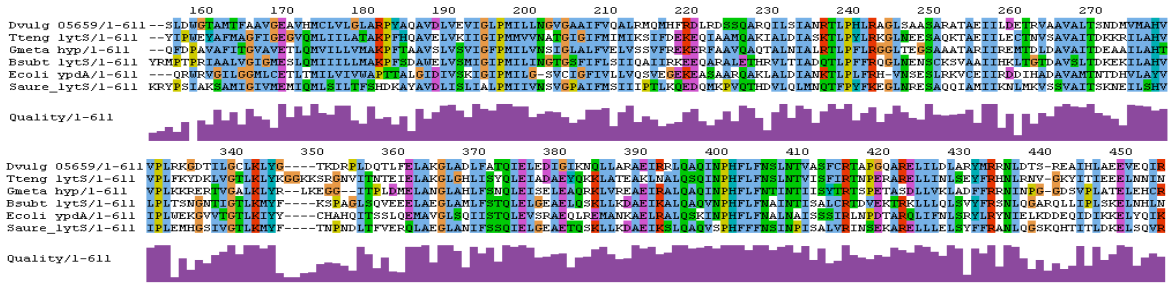
gradient elution to 80% buffer in 30 minutes. Acetonitrile was used as organic phase. The signals were detected with MRM positive mode of QTrap.

<u>*D. vulgaris* genetic analysis of two component systems.</u>

- HSK(s), TIGR ORF05707 and ORF5659 were used to test the knockout strategy devised earlier. Using the PCR to introduce 15bp overhangs the HSK *orf*(s) were cloned into the suicide plasmid pEX100T. The ligation free protocol was successful and reasonable transformation efficiency was achieved in *E.coli* XL10 Gold cells via heat shock. Approximately 10^3 colonies per nanogram of hybridized DNA were obtained. Colonies were successfully screened using blue/white selection and the inserts confirmed by colony PCR. Therefore the fist step in the knock out strategy has been tested.

- The ORF05659, was selected for knockout due to its annotation as LytS. This *D. vulgaris* LytS has homologs in several other bacteria (See figure below). In *S.aureus* the LytS knock out mutant showed a striking phenotype of altered cell structure clearly visualized via microscopy (Brunskill, et al,. J. Bacteriol. **178:**611-618). Given the homology, a *D. vulgaris* LytS knock out mutant might also be expected to have a similar phenotype.

- Conjugal transfer and electroporation into *D. vulgaris* was also attempted, although success of these methods has been limited thus far.



Future Work

- Set up CE-MS instrument.

- Determine antibiotic concentrations effective for *D. vulgaris* growth on plates.

- Complete a successful conjugation into *D. vulgaris*.

- Work out methods for cell lysis of *D. vulgaris* using French Press, and attempt to quantify differences between proteomes isolated using this method and sonication. This is in preparation for the actual completion of an oxygen stress experiment.

IV. Computational Core

LBNL – Arkin

In July, we completed the first phase of our quality assurance pipeline for the VIMSS Comparative Microbial Genomics Database. We have implemented an automated pipeline to check our own sequence and annotation information for each ORF against data available from the organizations hosting each genome.

We are still waiting for installation of our computer cluster; however, the vendor has made some progress in bringing our fileserver online. We purchased a 4 node web server to host our comparative genomics website, and it is currently undergoing testing offsite. We expect to install it within the next two weeks.

Much of our efforts have focused on bringing the comparative genomics website online by August 15th. We have integrated the Comparative Genomics Browser as well as a navigable metabolic map (for our target organisms) into the protein pages, and have started to work toward integrating our cis-regulatory annotations and predictions into the operon browser.

We have made some progress on two manuscripts describing our operon prediction tool. We are also planning to write a short note describing our comparative genomics web tools to accompany the public release. Finally, we are working with the FGC to do comparative analysis of ORNL microarray data with other species, as well as interpreting their data in light of our predicted operon and regulon structure for Shewanella.

LBNL – Olken

Contexts

We began considering how to model temporal aspects of environmental contexts (e.g., experimental protocols). Examples of such contexts include heat / cold shock experiments, oxygen shock experiments, etc. Discussions with experimentalists suggest that a very simple temporal model may be sufficient - i.e., all shocks are applied simultaneously at the beginning of the experiment, and the cell status is sampled at various time intervals after the shock(s) were applied.

Relational to Graph DB Mapping

We have decided to incorporate some aspect of federated DB early in the project, in particular to incorporate data from the BioSpice BioDB database. Hence, we have been concerned with schema mappings between relational and graph data models. We plan to (initially) use a naïve approach of mapping n-ary relations to a node plus n binary

relations (i.e., “edges”). Binary relations could be simply mapped to edges in the graph data model. See discussion of prototype development.

Prototype Development

We began development of a prototype to demonstrate the mapping from relational to graph data model. In this prototype we read the relational schema of the BioDB database via JDBC code, and then map the relational schema to a graph database schema.

The graph database schema is encoded as RDF (actually OWL). At present the graph representation (figure) only shows the table nodes, and edges corresponding to foreign keys, i.e., we suppress edges/nodes for attributes. This is done for the sake simplifying resulting diagrams. However, we do capture ordinary attributes. These are shown in the text window. The filtered schema is then converted to a DOT input file for ATT's Graphviz program, which is used to layout the schema as a PNG image and image map for use in a web browser. The user invokes a web browser on the schema image map. Mouse click (selection) events on the nodes invoke the presentation of detailed schema information for a particular node (table) in an adjoining browser frame. At present an alpha version of the schema browser works. Development of an instance browser is in progress. This program architecture is admittedly clumsy, but has shortened code development. HTML page generation for nodal schema pages is presently in Java – but we hope to convert to standard XML page generation, followed by XSLT and CSS style sheets for formatting.

Navigational API

We are currently working on a draft of a Java-based navigational API for accessing graphs from Java application programs. This programming interface is intended to insulate application programs from details of the data exchange and database access interface. Initially, this will be used with canned queries (or browsing), to be followed later by use with more sophisticated ad-hoc queries.

Collaborations

We have been contacted by James Ireland of Avalon Pharmaceutical Corporation, who is pursuing similar work on graph representations for pathway representation and searching, and also for LIMS (laboratory information management systems) applications, i.e., protocol representation and provenance recording and querying. We hope to be able to cooperate on code development and hope to meet with him in early August. We will meet with Meral Ozsoyoglu at the IEEE CS Bioinformatics meeting in mid-August and plan to discuss data modeling and query language issues. We posted a programmer position in the Scientific Data Management Group, funded by the Synnechococcus (Sandia) GTL project. The position has been advertised at the ISCB (Int'l. Society for Computational Biology) job board and the DBWORLD mailing list. We plan to interview candidates in August.

Plans for August

1. Prepare and give progress report to VIMSS GTL retreat in SF on Aug. 1-3, 2003.
2. Continue development of RDF/web based schema browser prototype.
3. Continue development of RDF/web based relational DB instance browser.
4. Complete biopathways chapter for DOE Computational Biology Primer.
5. Complete slides for Graph Data Management tutorial for IEEE CS Bioinformatics Meeting.
6. Present Graph Data Management tutorial for IEEE CS Bioinformatics Meeting.
7. Complete report on NLM Workshop on Data Management.
8. Complete Navigational API specification.
9. Code Navigational API.
10. Begin efforts to develop a common biopathway data interchange format with other members of Arkin Lab, the Ozsoyoglu's at Case Western Reserve Univ, Peter Mork (Univ. of Washington) and biopax.org.

V. Project Management

Project Schedule

The GTL project schedule is undergoing updates and will be posted to the VIMSS Discussion Board on a regular basis. Any updates/comments/revisions should be sent to Nancy Slater via email (naslater@lbl.gov).

GTL Project Meetings

- The next GTL Steering Committee Meeting will be held on September 3, 2003.

- The 2003 VIMSS/GTL Annual Retreat was held on August 1-3 at the Hotel Nikko in San Francisco. The posters and presentations are available on the Members Only area of the VIMSS website. The Retreat meeting notes are posted to the Discussion Board.